Obfuscated Gradients in GNN-Based NIDS Against Structural Adversarial Attacks on KDD Cup 99

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do obfuscated gradients in GNN-based NIDS models compare to other gradient masking techniques in terms of their robustness against structural adversarial attacks on the KDD Cup 99 dataset, as. The integration of machine learning (ML) algorithms into Internet of Things (IoT) applications has introduced significant advantages alongside vulnerabilities to adversarial attacks, especially within IoT-based intrusion detection systems (IDS). While theoretical adversarial. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Targeted Adversarial Traffic Generation : Black-box Approach to Evade Intrusion Detection Systems in IoT Networks. Research question: How do obfuscated gradients in GNN-based NIDS models compare to other gradient masking techniques in terms of their robustness against structural adversarial attacks on the KDD Cup 99 dataset, as measured by AUC-ROC scores?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

13 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Thirteen (13) sub-detectors are trained in parallel, with each specializing in a single feature.	×	0.01
The total number of features used in the study is thirteen.	×	0.03
Detection rate (Recall) is used as the primary metric to evaluate each sub-detector's ability to identify adversarial in	×	0.04
Reliability weights for sub-detectors are computed by normalizing detection rate values obtained from an independent eva	×	0.03
The defense architecture employs Bayesian fusion and Dempster-Shafer combination as ensemble fusion techniques.	×	0.02
The combined detector is positioned ahead of the NIDS to intercept and filter adversarial traffic before it reaches the	×	0.06
The defense architecture processes each instance by splitting it into 13 feature values.	×	0.03
In the threat scenario, the defender's NIDS operates as a flow-based system.	×	0.02
The attacker in the threat scenario operates under a black-box setting with no direct access to the IDS model's architec	×	0.06
The attacker constructs a substitute dataset that mimics the network traffic characteristics of the target IoT environme	×	0.04

References

- http://arxiv.org/abs/2211.10062v1
- http://arxiv.org/abs/2603.23438v1
- http://arxiv.org/abs/2512.10637v2